Damit rechnet doch keiner

Statistik Viele Studien sind unbrauchbar, weil Wissenschaftler schlecht planen und falsch rechnen, besagt eine neue Analyse - mit fatalen Folgen für das Ansehen der Forschung.

Viel vernichtender kann das Urteil für eine Disziplin nicht ausfallen. Was Neurowissenschaftler an Ergebnissen publizieren, sei äußerst zweifelhaft, schreibt ein Autorenteam um die Psychologin Katherine Button von der britischen University of Bristol in einer neuen Übersichtsstudie.

In vielen Arbeiten, so das Ergebnis der Auswertung von 49 Meta- und 730 Einzelstudien, würden die Autoren elementare Regeln der Statistik missachten. Die Methoden seien unzulässig, die Stichproben zu klein: Im Schnitt hätten die Studien nur eine statistische Teststärke von 21 Prozent. Bei Studien, in denen Gehirne per funktioneller Magnetresonanztomografie (fMRT) beobachtet werden, seien es sogar nur 8 Prozent. (Nature Reviews Neuroscience, Link: doi.org/k9d).

Im Bann der Signifikanz

Was bedeutet: Die Studien sind so schlecht konzipiert, dass sie nur eine 21- beziehungsweise 8-prozentige Chance haben, einen Effekt aufzuspüren, der tatsächlich existiert. Umgekehrt steigt die Gefahr, eine Wirkung auszumachen, wo in Wirklichkeit keine ist.

Selbst wenn sie einen Effekt richtig erkennen, sind schwache Studien anfälliger für den „Fluch des Siegers“. So nennen Statistiker das Phänomen, dass die erste Studie, die einen Effekt nachweist, ihn oft stark überschätzt. Weswegen Folgestudien die publizierten Ergebnisse häufig nicht bestätigen können.

Die Klage ist weder neu noch auf die Neurowissenschaften beschränkt: Viele Naturwissenschaftler vergessen im Labor, was sie im Grundstudium über Statistik gelernt haben. Sie arbeiten mit zu kleinen Stichproben. Sie stellen in ihren Berechnungen keine wohl durchdachte Hypothese auf den Prüfstand. Stattdessen melken sie ihre Daten so lange mit immer neuen Methoden, bis sie ein Ergebnis haben, dem sie unzulässigerweise Signifikanz bescheinigen. Denn eine Arbeit, die einen Effekt nachweist, hat bessere Chancen auf eine Publikation – und der Druck zu veröffentlichen ist enorm.

Um das Wissen vieler Gutachter und Herausgeber von wissenschaftlichen Fachmagazinen, so die Kritik, ist es nicht besser bestellt; munter winken sie Arbeiten durch, deren Verfasser sich mit unzulässigen Methoden das Gütesiegel signifikant ausgerechnet haben.

Für eine Publikation rechnen Forscher sich die Daten schön.

Schon 2005 ging der Mediziner John Ioannidis – Co-Autor der neuen Studie – mit der Fachwelt hart ins Gericht. Unter dem provokanten Titel „Why Most Published Research Findings Are False“ forderte er dazu auf, sich von der blinden Jagd nach Signifikanz zu verabschieden und wieder mehr Wert auf methodische Qualität zu legen (Plos Medicine, Link: doi.org/chhf6b).

Und der Zellbiologe David Vaux von der University of Melbourne sah sich erst im Dezember 2012 genötigt, seine Fachkollegen im Fachmagazin Nature an ein paar statistische Grundbegriffe und Gesetze zu erinnern – weil schlampig geplante und durchgeführte Studien dort wie in anderen Fachmagazinen inzwischen an der Tagesordnung seien (Link: doi.org/k8h).

Der Professor zürnt

„Die Gemeinde, die sagt: ‚Das ist alles Schwachsinn, was wir gemacht haben‘, wird zwar immer größer“, meint der Statistiker Walter Krämer von der Technischen Universität Dortmund. „Allerdings sind die Leute, die den Schwachsinn weitertreiben, weiterhin in der Mehrheit.“ In seinen Vorlesungen versucht er, einer in seinen Augen übertriebenen Signifikanzgläubigkeit ein kritisches Denken entgegenzusetzen: „Ich hab meinen Studenten gesagt: Das Wort signifikant möchte ich nicht mehr hören. Ich möchte vielmehr Begründungen haben, weshalb das, was sie gefunden haben, nicht auf dem Zufall beruht.“

Um zu erklären, wie irreführend Signifikanz sei, benutzt Krämer gern ein Beispiel aus der Finanzwelt: „An Börsentagen, die geteilt durch 7 den Rest 1 ergeben – also am 1., am 8., 15., am 22. und 29. eines Monats –, gibt es signifikant höhere Renditen an den deutschen

Tolle Technik schützt vor miesen Daten nicht.

Aktienmärkten.“ Und schiebt direkt nach: „So ein Schwachsinn.“

Den vermeintlichen Zusammenhang entdeckte Krämer, als er Hunderte statistische Berechnungen mit den gleichen Börsendaten durchführte – bis eine einen Zusammenhang aufzeigte.

Überstrapazierte Daten

Als signifikant gelten Erkenntnisse, wenn mit einer hohen Wahrscheinlichkeit ausgeschlossen werden kann, dass hier nur der Zufall am Werk war; als kritische Schwelle gilt oft ein Signifikanzniveau von 95 Prozent. Was im Umkehrschluss bedeutet, dass eine 5-prozentige Chance besteht, einen Effekt aufzudecken, den es gar nicht gibt.

„Wenn kein System hinter einem Effekt steckt, also alles nur Zufall ist“, sagt Krämer, „und eine wissenschaftliche Prozedur in fünf Prozent dieser Fälle trotzdem einen systematischen Effekt anzeigt, dann werden Sie bei hundert Experimenten in fünf etwas Signifikantes finden.“ Die Börsentagsformel war ein Artefakt, ein Teil der zufällig funktionierenden fünf Prozent.

Gegen die Versuchung, mit neuen Berechnungen an seinen Daten herumzuwerkeln, empfehlen die Autoren um Katherine Button die Datenbank des Open Science Framework (OSF). Hier können Forscher ihre Ergebnisse hinterlegen und melden, welche Hypothese sie mit welchen Methoden prüfen. Das OSF ist nur eines von mehreren Netzwerken, die sich der Forschungsqualität verschrieben haben und Wissenschaftler dazu ermuntern, ihre Daten offenzulegen. Projekte wie das Dataverse Network oder OpenfMRI bieten eine Bühne für verschiedene Disziplinen.

Die Methodenkritiker sorgen sich nicht nur um das Ansehen der Wissenschaft und der Fachjournale. „Geringe Teststärke hat auch einen ethischen Aspekt“, schreibt Button. Für schlechte Studien würden Mittel verschwendet, vor allem aber Versuchstiere und Probanden unnötig in Mitleidenschaft gezogen – ganz zu schweigen von den Patienten, deren Behandlung auf verzerrten Studienergebnissen beruht.

„Schlimmstenfalls kann Statistik töten“, sagt Gerd Antes, Direktor des Deutschen Cochrane-Zentrums, das sich der Verbreitung evidenzbasierter, also wissenschaftlich valider Medizin verschrieben hat. Er empfiehlt die Richtlinien des Equator-Netzwerks (Link: www.equator-network.org), das Checklisten und andere Hilfsmittel für valide Publikationen erarbeitet.

Besser doppelblind?

David Vaux plädiert dafür, dass junge Forscher vor ihren ersten Experimenten das Denken in statistischen Zusammenhängen lernen. Um die Qualität von Veröffentlichungen zu verbessern, sollten Fachjournale jeden Artikel doppelblind begutachten lassen. Bei diesem Verfahren kennt der Autor nicht den Namen des Gutachters und umgekehrt. „Derzeit beeinflusst nicht der Inhalt eines Papers die Veröffentlichung, sondern wer die Autoren sind. Das System ist zu vorurteilsbeladen, um schlechte Publikationen herauszufischen.“ Walter Krämer zweifelt: „Selbst wenn ein Review-Prozess doppelblind ist: Nach drei Zeilen wissen Sie, wer’s geschrieben hat. Sie kennen doch die Gemeinde.“ Auch eine Sprecherin des Magazins Nature hält doppelblinde Gutachten für wenig sinnvoll, die Redaktion setze auf eigene Statistiker, die jeden Aufsatz prüfen.

Nicht gut genug, so Vaux, den neben seiner Forscherethik auch gesunder Eigennutz umtreibt: „Schlampen Autoren, Gutachter und Redakteure, erhöht das die Chance, beim Lesen Zeit zu verschwenden.“

LUKAS SCHÜRMANN, GEORG DAHM